PARSEME Short term scientific mission
Activity report

Eric Wehrli
University of Geneva
28/4/2015

1. Purpose of the STSM

The main goal of this mission was to initiate a collaboration between Aline Villavicencio's research group at the Federal University of Rio Grande do Sul (UFRGS) and Eric Wehrli's group at the University of Geneva in the domain of MWEs in both Portuguese and English, with a longer term objective of creating a bilingual database of MWEs available on-line.

2. The mission took place last in March and April 2015

Besides the usual academic exchanges, public conference, discussions with colleagues and with graduate students, the main achievements carried out during the STSM (including preparation over several months) are:

- the development of a preliminary version of a syntactic parser for Portuguese, based on the Fips parser, that will be used for the extraction of collocations.
- elaboration of a lexical database for Portuguese, with a morphological engine to generate all the inflected forms for verbs, nouns and adjectives (about 300,000 orthographical forms corresponding to nearly 20,000 lexemes).
- UFRGS will create a bilingual database (Portuguese, English) of newspaper articles, which -- along with other aligned bilingual corpora, such as europarl -- will serve as basis for MWE extraction.

The invited talk was attended by a large audience from Computer Science, Linguistics and other departments at UFRGS, and was an important means to disseminate work on MWEs and collocations, which raised awareness and attracted considerable interest.

3. Plans :

- We intend to submit a proposal for a communication at the next PARSEME meeting in Iasi, Romania, describing our research plans and early results.
- UFRGS will extract MWEs from the bilingual corpora (see above) using the mwetoolkit, while UNIGE will extract collocations from the same corpora by means of the Fips collocation extraction tool (based on the syntactic parser mentioned above).